IMG: Calibrating Diffusion Models via Implicit Multimodal Guidance

Supplementary Material

A. Implementation Details

A.1. Baselines and Models.

Our experiments are based on two diffusion models: SDXL [57], a widely adopted base diffusion model for alignment tasks, and FLUX.1 [dev] (FLUX) [37], a recent state-of-the-art flow-matching-based diffusion transformer. To compare with finetuning-based methods, we use the top-performing finetuned variant of SDXL, SDXL-DPO, which applies the Diffusion-DPO [67] technique, demonstrating the superiority of IMG and its compatibility with finetuning-based methods. For comparison with editing-based methods, we adopt the leading SLD as our baseline to highlight the advantages of IMG in visual comprehension and aesthetic quality. We further compare IMG with leading compositional generation methods, ELLA [26] and CoMat [33], to evaluate the compositional generation capabilities. For MLLM, we finetune LLaVA 1.5-13b [42] on the Instruct-Pix2Pix dataset [6] for 1 epoch, using the finetuning task format shown in Fig. 11, and extract features from the last hidden layer for guidance. We utilize the IP-Adapter [72, 78], trained on SDXL and FLUX, to enable image prompts and extract image features. The Implicit Aligner takes both MLLM and image features as input and is implemented as a stack of 4 cross-attention layers and 2 linear layers. A detailed illustrative diagram of Implicit Aligner is shown in Fig. 9, accompanied by its execution pseudo code in Fig. 10.

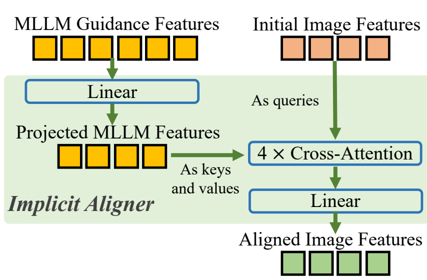

Figure 9. **Detailed architecture of Implicit Aligner.** Our Implicit Aligner contains 4 cross-attention layers and 2 linear layers. The number of color cubes here represents the token dimensions rather than the number of tokens.

A.2. Datasets and Benchmarks.

For Implicit Aligner training, we use the same Pick-a-Pic training set [36] as Diffusion-DPO [67], which consists of 851K pairs of preferred and unpreferred images generated under specific prompts. The preference labels are annotated by human observers. To determine the optimal training scheme and hyperparameters, we conduct ablation studies by evaluating the average Pick Score [36] across generated images using 500 unique prompts from the Pick-a-Pic test set. The Pick Score is a caption-aware preference scoring model trained on Pick-a-Pic. For evaluation, we report Human Preference Scores v2 (HPS v2) across generated images on the Human Preference Datasets v2 (HPD v2) test set [71], which includes 3,400 prompts across five categories, as well as the Parpi-Prompts [79], a diverse dataset of 1,632 prompts ranging from brief concepts to complex sentences. HPS v2 is a caption-aware preference scoring model trained on HPD v2. We also report results on the T2I-CompBench [28], which contains 1800 test prompts to validate compositional image generation capabilities. For each test in user studies, 33 evaluators were asked to do an A-B test on 30 random image pairs generated by the base model and IMG with the same prompt. Each unique pair was assessed by 3 evaluators, and only fully consistent votes were used to compute the final win rates. For MLLM finetuning, we extract triplets of {Original Image, Edited Prompt, Edit Instruction} from the CLIP-filtered Instruct-Pix2Pix dataset [6], which contains 313K samples.

```
# Input: feat_i [b, s_1, d_1]: initial image features
# Input: feat_m [b, s_2, d_2]: MLLM guidance features
# Output: feat_a [b, s_1, d_1]: aligned image features
# Step 1: project feat_m to dimension d_1
feat_m_proj = Linear(feat_m) # [b, s_2, d_1]
# Step 2: cross-attention between feat_i and feat_m_proj
atten = Cross-attention(q=feat_i, k,v=feat_m_proj)
# Step 3: process atten via a Linear layer as feat_a
feat_a = Linear(feat_a) # [b, s_1, d_1]
```

Figure 10. **Pseudo code of Implicit Aligner.** Our Implict Aligner (1) projects MLLM features to the same dimension as image features; (2) conducts cross-attention between initial image features and projected MLLM features; and (3) processes attention outputs with a linear layer as aligned image features.

A.3. MLLM Finetuning.

To customize a pretrained MLLM as a misalignment detector, we finetune LLaVA 1.5-13b [42] on the Instruct-Pix2Pix dataset [6] for 1 epoch. We use training triplets consisting of original images I_0, edited prompts T_1, and edit instructions T_E. While I_0 and T_1 are fed into the MLLM as inputs, we prompt the model to describe the alignment by asking questions such as, 'How can the <Original Image> match the intended prompt: <Edited Prompt>?', and supervise the model's outputs against T_E (see Fig. 11). To prevent overfitting, we randomly select one of 100 different misalignment detection questions for each sample. The

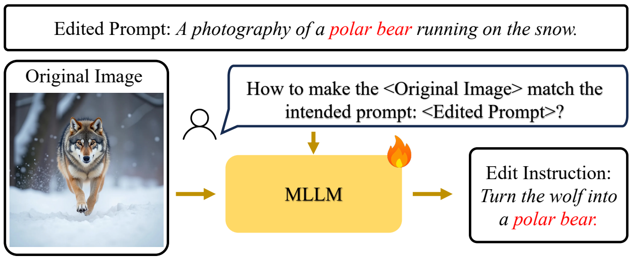

Figure 11. **MLLM finetuning on instruction-based image data.** We conduct finetuning on {Original Image, Edited Prompt, Edit Instruction} triplets from image editing datasets [6] to enhance MLLM's comprehension on prompt-image misalignments.

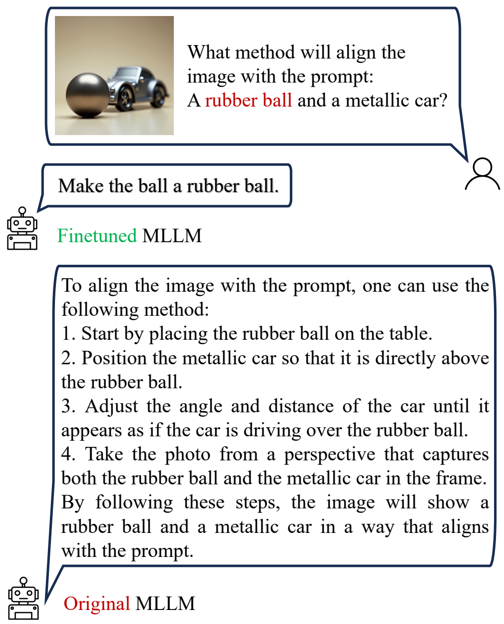

Figure 12. Text response comparison of the original MLLM and our finetuned MLLM. The original MLLM primarily outlines an image generation process based on the prompt, while our finetuned MLLM emphasizes aligning the input image with the provided prompt, showcasing its misalignment detection capability.

fine-tuning hyperparameters follow the standard configurations in [42]. In Fig. 12, we compare the text responses of the original MLLM and our fine-tuned MLLM. The original MLLM primarily outlines an image generation process based on the prompt, while our finetuned MLLM emphasizes aligning the input image with the provided prompt, showcasing its misalignment detection capability.

A.4. IMG Training and Evaluation.

Our Implicit Aligner is trained on the Pick-a-Pic training set [36] for 100K iterations with 8 A100 GPUs and a batch size of 8. We use the AdamW [45] optimizer with a constant learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The ratio parameter in Eq. 7 is set to 1. The reference model updating step k in Sec. 3.4 is set to 10. The training process takes about 10-15 hours. For evaluation, we set classifier-free guidance [24] to 7.5 for SDXL and SDXL-DPO, and 3.5 for FLUX. Sampling steps are set to 50 for SDXL and SDXL-DPO, and 30 for FLUX. The MLLM in IMG consume about 4% additional inference time and 15G (Qwen-VL-7B) - 25G (LLaVA-13B) GPU memory.

B. Objective Derivation

This section presents the detailed derivation of our proposed Iteratively Updated Preference Objective L in Sec. 3.4, which is a combination of a basic objective L_{base} and a preference objective L_{pref}. To enhance generality and clarity, we substitute the $\boldsymbol{c}_I^w$ and $\boldsymbol{c}_I^l$ in Sec. 3.4 with more general forms, $\boldsymbol{x}_w$ and $\boldsymbol{x}_l$. These denote the preferred and non-preferred outputs of a regression model f_θ (the Implicit Aligner in IMG), under a given condition $\boldsymbol{c}$. In essence, the training procedure operates on triplets $\{\boldsymbol{c}, \boldsymbol{x}_w, \boldsymbol{x}_l\}$.

B.1. Basic Objective

The primary goal of f_θ is to predict the preferred sample $\boldsymbol{x}_w$, given the condition $\boldsymbol{c}$, as formalized in Eq. 4:

$$L_{\text{base}} = \mathbb{E}_{\boldsymbol{c},\boldsymbol{x}_w} ||\boldsymbol{x}_w - f_\theta(\boldsymbol{c})||_2^2. \tag{8}$$

Minimizing the above Mean Square Error(MSE) is a well-established approach, equivalent to performing maximum likelihood estimation (MLE) in regression settings [39, 52, 63]. Within this framework, $f_\theta(\boldsymbol{c})$ predicts the mean of a noisy distribution, which is assumed to follow a Gaussian distribution with constant variance σI, consistent with the probabilistic interpretation [48]:

$$p_\theta(\boldsymbol{x}_w|\boldsymbol{c}) = N(\boldsymbol{x}_w|f_\theta(\boldsymbol{c}), \sigma \mathrm{I}). \tag{9}$$

The MSE in Eq. 8 equals the negative log-likelihood (NLL) of $p_\theta(\boldsymbol{x}_w|\boldsymbol{c})$ [52]. Consequently, training the regression model f_θ using MSE implicitly enables it to approximate the conditional data distribution $p_{\text{data}}(\boldsymbol{x}_w|\boldsymbol{c})$.

B.2. Preference Objective

Besides the basics, we also draw inspiration from direct preference optimization (DPO) [67] and self-play finetuning (SPIN) [80] to enhance alignment. These preference learning techniques adhere to a common RLHF principle [60]: optimize the conditional distribution $p_\theta(\boldsymbol{x}|\boldsymbol{c})$ to maximize a latent reward model $r(\boldsymbol{c}, \boldsymbol{x})$, while regularizing the KL-divergence from a reference distribution p_{ref}:

$$\max_{p_\theta} \mathbb{E}_{\boldsymbol{c},\boldsymbol{x}}[r(\boldsymbol{c}, \boldsymbol{x})] - \eta \mathrm{KL}\left(p_\theta(\boldsymbol{x}|\boldsymbol{c})||p_{\text{ref}}(\boldsymbol{x}|\boldsymbol{c})\right). \tag{10}$$

Here p_θ and p_{ref} are prediction distributions of f_θ and f_{ref}, respectively, where f_{ref} is a copy of f_θ from an earlier training iteration, as defined in Eq. 9. The hyperparameter η controls the strength of the regularization.

As demonstrated in [60], the unique global optimal solution of $p_\theta(\boldsymbol{x}|\boldsymbol{c})$ in Eq. 10 is expressed as:

$$p_\theta(\boldsymbol{x}|\boldsymbol{c}) = p_{\text{ref}}(\boldsymbol{x}|\boldsymbol{c}) \exp\left(r(\boldsymbol{c}, \boldsymbol{x})/\eta\right)/Z(\boldsymbol{c}), \tag{11}$$

where $Z(\boldsymbol{c}) = \sum_{\boldsymbol{x}_0} p_{\text{ref}}(\boldsymbol{x}_0|\boldsymbol{c}) \exp\left(r(\boldsymbol{c}, \boldsymbol{x}_0)/\eta\right)$ is the partition function. The reward model is reformulated as:

$$r(\boldsymbol{c}, \boldsymbol{x}) = \eta \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}|\boldsymbol{c})} + \eta \log Z(\boldsymbol{c}). \tag{12}$$

From the perspective of integral probability metric (IPM) [51], DPO [67] maximizes the reward gap between preferred and non-preferred data distributions, while SPIN [80] maximizes the reward gap between preferred data distribution and reference data distribution, *i.e.*, $\boldsymbol{x}_{\text{ref}} = f_{\text{ref}}(\boldsymbol{c}) \sim p_{\text{ref}}(\boldsymbol{x}|\boldsymbol{c})$. As introduced in Sec. 3.4, we establish a combined objective of DPO and SPIN:

$$\begin{aligned} \max_r \mathrm{E}_{\boldsymbol{c},\boldsymbol{x}_w,\boldsymbol{x}_l,\boldsymbol{x}_{\text{ref}}} [&\underbrace{r(\boldsymbol{c}, \boldsymbol{x}_w) - r(\boldsymbol{c}, \boldsymbol{x}_l)}_{\text{DPO}} \\ &+ \mu(\underbrace{r(\boldsymbol{c}, \boldsymbol{x}_w) - r(\boldsymbol{c}, \boldsymbol{x}_{\text{ref}})}_{\text{SPIN}})], \end{aligned} \tag{13}$$

where μ is a hyperparameter that controls the trade-off. As demonstrated by [8], a more general form of the optimization problem in Eq. 13 is:

$$\begin{aligned} \min_r \mathrm{E}_{\boldsymbol{c},\boldsymbol{x}_w,\boldsymbol{x}_l,\boldsymbol{x}_{\text{ref}}} [&\ell(r(\boldsymbol{c}, \boldsymbol{x}_w) - r(\boldsymbol{c}, \boldsymbol{x}_l) \\ &+ \mu(r(\boldsymbol{c}, \boldsymbol{x}_w) - r(\boldsymbol{c}, \boldsymbol{x}_{\text{ref}})))], \end{aligned} \tag{14}$$

where ℓ represents any monotonically decreasing convex loss function. Eq. 13 can be viewed as the maximization version of Eq. 14, where $l(a) = -a$. However, using such a linear loss function leads to an unbounded objective value, which may cause undesirable negative infinite values of $r(\boldsymbol{c}, \boldsymbol{x}_l)$ and $r(\boldsymbol{c}, \boldsymbol{x}_{\text{ref}})$ during continuous training. To address this issue, we adopt a logistic loss function as suggested by [67, 80]:

$$l(a) := -\log \text{sigmoid}(a) = \log(1 + \exp(-a)), \tag{15}$$

which is non-negative, smooth, and exhibits an exponentially decaying tail as $a \to \infty$. The logistic loss function helps prevent the excessive growth of the reward value r, ensuring a stable training process.

By substituting the reward model r in Eq. 14 with Eq. 12 and empirically setting η and μ to 1, we obtain the final preference objective as follows:

$$\begin{aligned} L_{\text{pref}} = \mathbb{E}_{\boldsymbol{c},\boldsymbol{x}_w,\boldsymbol{x}_l,\boldsymbol{x}_{\text{ref}}} \Bigg[\ell \Bigg(&\log \frac{p_\theta(\boldsymbol{x}_w|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{x}_w|\mathbf{c})} - \log \frac{p_\theta(\boldsymbol{x}_l|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{x}_l|\mathbf{c})} \\ &+ \log \frac{p_\theta(\boldsymbol{x}_w|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{x}_w|\mathbf{c})} - \log \frac{p_\theta(\boldsymbol{x}_{\text{ref}}|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{x}_{\text{ref}}|\mathbf{c})} \Bigg) \Bigg], \end{aligned} \tag{16}$$

which aligns with Eq. 5. Using the equivalence between MSE and NLL under the Gaussian prior, as discussed in Appendix B.1, we obtain a simplified version of L_{pref} for implementation as follows:

$$\begin{aligned} L_{\text{pref}} = \mathbb{E}_{\boldsymbol{c},\boldsymbol{x}_w,\boldsymbol{x}_l} [&\ell(-[2(||\boldsymbol{x}_w - f_\theta(\boldsymbol{c})||_2^2 - ||\boldsymbol{x}_w - f_{\text{ref}}(\boldsymbol{c})||_2^2) \\ &- (||\boldsymbol{x}_l - f_\theta(\boldsymbol{c})||_2^2 - ||\boldsymbol{x}_l - f_{\text{ref}}(\boldsymbol{c})||_2^2) \\ &- ||f_{\text{ref}}(\boldsymbol{c}) - f_\theta(\boldsymbol{c})||_2^2])], \end{aligned} \tag{17}$$

which is consistent with Eq. 6. As discussed in Sec. 3.4, the reference model f_{ref} is iteratively updated. Specifically, we first randomly initialize f_{ref} and later iteratively copy f_θ to f_{ref} whenever f_θ outperforms f_{ref}. In practice, we execute the substitution when $f_\theta(c)$ is closer to $\boldsymbol{x}_w$ than $f_{\text{ref}}(c)$ for k consecutive iterations, *i.e.*,

$$||\boldsymbol{x}_w - f_\theta(\boldsymbol{c})||_2^2 < ||\boldsymbol{x}_w - f_{\text{ref}}(\boldsymbol{c})||_2^2. \tag{18}$$

To summarize, The final Iteratively Updated Preference Objective is a combination of L_{base} and L_{pref}, weighted by a ratio parameter λ:

$$L = L_{\text{base}} + \lambda L_{\text{pref}}. \tag{19}$$

C. Additional Quantitative Results

In Tab. 6, we present additional quantitative results on GenEval [17] and DPGBench [26]. IMG shows consistent improvements across two benchmarks.

Model	GenEval↑	DPGBench↑
SDXL-DPO	0.59	76.81
SDXL-DPO + IMG (Ours)	**0.61**	**78.72**
FLUX	0.68	80.60
FLUX + IMG (Ours)	**0.70**	**82.77**

Table 6. Results on GenEval [17] and DPGBench [26].

D. Additional Qualitative Results

In Fig. 13, we compare IMG with leading MLLM-based image editing methods [15, 30]. IMG showcases better alignment performance and visual quality.

In Fig. 14 and Fig. 15, we present additional qualitative results to show the superior prompt adherence and aesthetic quality achieved by integrating IMG with various models.

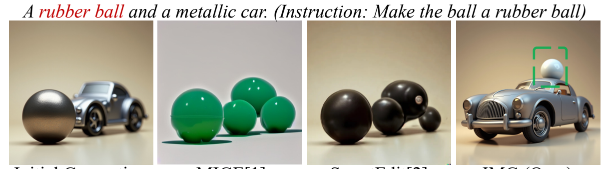

Figure 13. Comparison between MLLM-based editing and IMG.

Figure 14. Additional qualitative results by integrating IMG with FLUX.

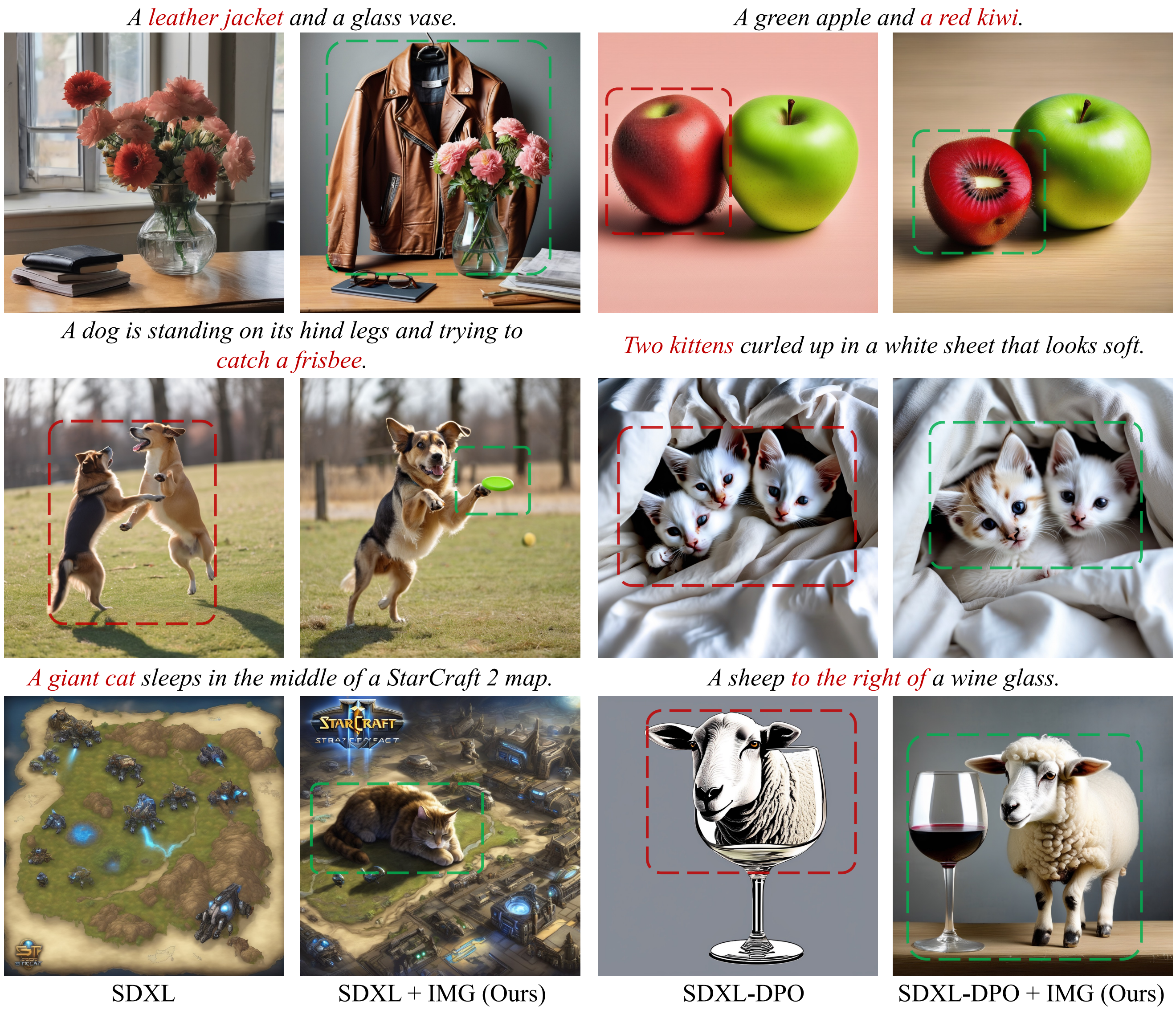

Figure 15. Additional qualitative results by integrating IMG with SDXL and SDXL-DPO.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 2, 3, 7

[2] Stability AI. Stable diffusion 3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024. 7, 8

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 3

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7

[5] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2023. 2, 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3, 4, 5, 1, 2

[7] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv:2403.04692*, 2024. 7, 8

[8] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *ICML*, 2024. 2, 5, 3

[9] Mang Tik Chiu, Yuqian Zhou, Lingzhi Zhang, Zhe Lin, Connelly Barnes, Sohrab Amirghodsi, Eli Shechtman, and Humphrey Shi. Brush2prompt: Contextual prompt generator for object inpainting. In *CVPR*, 2024. 3

[10] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv:2309.17400*, 2023. 2, 3

[11] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv:2309.15807*, 2023. 2

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 2, 3

[13] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *NeurIPS*, 2024. 2, 3

[14] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. In *NeurIPS*, 2024. 3

[15] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *ICLR*, 2024. 3

[16] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 7

[17] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023. 3

[18] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair diffusion: A comprehensive multimodal object-level image editor. In *CVPR*, 2024. 3

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[20] Jiayi Guo, Chaoqun Du, Jiangshan Wang, Huijuan Huang, Pengfei Wan, and Gao Huang. Assessing a single image in reference-guided image synthesis. In *AAAI*, 2022. 3

[21] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Humphrey Shi, Gao Huang, and Shiji Song. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, 2023. 3

[22] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *CVPR*, 2024. 3

[23] Jiayi Guo, Junhao Zhao, Chaoqun Du, Yulin Wang, Chunjiang Ge, Zanlin Ni, Shiji Song, Humphrey Shi, and Gao Huang. Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. In *CVPR*, 2025. 3

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021. 2

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[26] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv:2403.05135*, 2024. 5, 6, 7, 1, 3

[27] Jiannan Huang, Jun Hao Liew, Hanshu Yan, Yuyang Yin, Yao Zhao, Humphrey Shi, and Yunchao Wei. Classdiffusion: More aligned personalization tuning with explicit class guidance. In *ICLR*, 2025. 3

[28] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023. 5, 1

[29] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv:2302.09778*, 2023. 3

[30] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui

Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, 2024. 3
[31] Arman Isajanyan, Artur Shatveryan, David Kocharian, Zhangyang Wang, and Humphrey Shi. Social reward: Evaluating and enhancing generative AI through million-user feedback from an online creative community. In *ICLR*, 2024. 3
[32] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *CVPR*, 2024. 3
[33] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. In *NeurIPS*, 2024. 5, 6, 1
[34] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, 2024. 3
[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2015. 5
[36] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. 3, 4, 5, 1, 2
[37] Black Forest Labs. Flux. https://blackforestlabs.ai/, 2024. 1, 2, 3, 5, 6, 8
[38] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv:2402.17245*, 2024. 7, 8
[39] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *CVPR*, 2021. 5, 2
[40] Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. In *ACL*, 2021. 3
[41] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *TMLR*, 2024. 3
[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 3, 5, 1, 2
[43] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2022. 3
[44] Zeyu Liu, Zanlin Ni, Yeguo Hua, Xin Deng, Xiao Ma, Cheng Zhong, and Gao Huang. Coda: Repurposing continuous vaes for discrete tokenization. *arXiv preprint arXiv:2503.17760*, 2025. 3
[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5, 2
[46] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 3
[47] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *CVPR*, 2023. 3
[48] Peter McCullagh. *Generalized linear models*. Routledge, 2019. 2
[49] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *NeurIPS*, 2023. 2, 3
[50] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv:2302.08453*, 2023. 3, 6
[51] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 1997. 3
[52] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *ICNN*. IEEE, 1994. 2
[53] OpenAI. Openai o1. https://openai.com/o1/, 2024. 2
[54] OpenAI. Openai o3-mini. https://openai.com/index/openai-o3-mini/, 2024. 2
[55] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. *arXiv:2405.16537*, 2024. 3
[56] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *ICLR*, 2024. 3
[57] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2023. 2, 5, 6, 1
[58] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv:2305.11147*, 2023. 3
[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3
[60] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2024. 2, 3, 5
[61] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 3
[62] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 3
[63] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *CVPR*, 2022. 5, 2

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 3

[66] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. In *ECCV*, 2024. 3

[67] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024. 2, 3, 5, 6, 1

[68] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv:2404.02733*, 2024. 5

[69] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. In *NeurIPS*, 2024. 3

[70] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *CVPR*, 2024. 2, 3, 4, 5, 6, 7

[71] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv:2306.09341*, 2023. 2, 5, 6, 1

[72] XLabs-AI. X-flux. `https://github.com/XLabs-AI/x-flux`, 2024. 3, 5, 1

[73] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv:2304.05977*, 2023. 3

[74] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile Diffusion: Text, images and variations all in one diffusion model. *arXiv:2211.08332*, 2022. 2

[75] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *CVPR*, 2024. 3

[76] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *CVPR*, 2024. 3

[77] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *ICML*, 2024. 3

[78] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023. 3, 5, 6, 1

[79] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022. 5, 6, 7, 1

[80] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *arXiv:2402.10210*, 2024. 2, 3, 5

[81] Gong Zhang, Kihyuk Sohn, Meera Hahn, Humphrey Shi, and Irfan Essa. Finestyle: Fine-grained controllable style personalization for text-to-image models. In *NeurIPS*, 2024. 3

[82] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[83] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 3

[84] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. In *ICML*, 2024. 3